Spell Checker for Hindi

Amit Sharma (10080) and Pulkit Jain (10543)

CS365 Course Project under Dr. Amitabha Mukherjee

INTRODUCTION

Spell checking applications are important part of several fundamental applications such as editors and search engines. The task of spell checking is vital in providing correct and quality information through text. The task of spell checking may be divided into identifying erroneous words in the text and replacing these with the correct intended words.

MOTIVATION

Spell checking has been researched into a great depth. While there are state of the art spell checker tools available for English language, not much work has been done in this field for Hindi. Various documents, novels, newspapers are typed in Hindi and there is a need for development of spell checker tools for Hindi. We, in this project aim at building a spell checker application for Hindi language.

RELATED WORK

Spelling errors arising in documents are broadly classified as non-word errors and real-word errors. Non-word errors are those errors where a word is spelled incorrectly. Ex: आंख for आँख (eye). Real word errors include those errors where the wrongly spelt word is a valid word in the language but is not the intended word in the language. Ex: misspelling गृह (home) with ग्रह (planet), सुत (son) with सूत (cotton).

Multiple approaches that have been developed to solve the non-word spell checking problem include N-Gram analysis and the dictionary lookup for identifying the errors and edit distance approach. Similarity keys, N-Gram based approach and rule based techniques are some of the approaches for predicting corrections to wrongly spelled words, as stated in [4]. Spell checking has also been done by modeling languages as finite state automaton and assigning weights to error corrections, as presented in [1]. However, the lexicon based approach for identifying the errors combined with the shortest edit distance approach is the most widely used combination for existing spell checker applications.

Correcting real-word errors requires building the context in which the word is introduced. Context based spell checking for correcting tags has been done in work presented in [2]. Several methods for building context sensitive spell checkers, including the N-Gram analysis have been mentioned in [3].

OUR APPROACH

We are primarily going to follow the work done in [3].

- We aim at first building a lexicon based spell checker by building a lookup dictionary from the available Hindi Corpus. Errors will be detected using dictionary lookup and corrections will be suggested on the basis on edit distances. Word frequency might be used for ranking words at same edit distances.

- Context based checking for real-word errors will be done using the N-Gram approach described in [3]. 3-grams, 4- grams or 5-grams constructed from the given text to be checked will be searched for in a set of N-grams constructed from available corpus. Frequencies of these N-grams will be used to suggest possible corrections for real world errors.

RESOURCES

- Hindi Corpus made available at http://www.cfilt.iitb.ac.in/hin_corp_unicode.tar will be used.
- Code made available at http://norvig.com/spell-correct.html will be used as a starting point.

REFERENCES

[1] Tommi Pirinen and Krister Linden. Finite-state spell-checking with weighted language and error models. Proceedings of LREC 2010 Workshop on Creation and use of basic lexical resources for less-resourced languages [2010]

[2] Francesco Bonchi, Ophir Frieder, Franco Maria Nardini, Fabrizio Silvestri and Hossein Vahabi. Interactive and Context-Aware Tag Spell Check and Correction [2012]

[3] Suzan Verberne. Context-sensitive spell checking based on word trigram probabilities [2002]

[4] Neha Gupta, Pratistha Mathur. Spell Checking Techniques In NLP: A Survey [2012]

[5] Peter Norvig. How to write a spelling corrector. http://norvig.com/spell-correct.html